Education and Information Technologies
https://doi.org/10.1007/s10639-025-13553-1

ChatGPT: A reliable assistant for the evaluation of students' written texts?

Arzu Atasoy[1] · **Saieed Moslemi Nezhad Arani**[2]

Received: 21 October 2024 / Accepted: 25 March 2025

Abstract

There is growing interest in the potential of Artificial Intelligence (AI) to assist in various educational tasks, including writing assessment. However, the comparative efficacy of human and AI-powered systems in this domain remains a subject of ongoing exploration. This study aimed to compare the accuracy of human raters (teachers and pre-service teachers) and AI systems (ChatGPT and trained ChatGPT) in classifying written texts. The study employed both chi-square tests and logistic regression analysis to examine the relationship between rater groups (human vs. machine) and the accuracy of text classification. Initial chi-square analyses suggested no significant differences in classification accuracy between human and AI raters. However, the logistic regression model revealed a significant relationship, with human raters demonstrating a higher rate of correct classification compared to their AI counterparts. The logistic model achieved an 81.3% success rate in predicting correct classifications. While AI systems show promise in automated text processing, human raters currently demonstrate superior accuracy in writing assessment tasks. These findings highlight the need for further research into the strengths and limitations of both human and AI-based approaches. The integration of AI in educational assessment should focus on complementing and supporting, rather than replacing, the expertise of human educators.

Keywords Assessing writing · ChatGPT · Rating texts · Turkish texts

✉ Arzu Atasoy
arzuatasoy@gantep.edu.tr

Saieed Moslemi Nezhad Arani
s.mosleminezhad@bam.ac.ir

[1] Department of Turkish and Social Sciences Education, Education Faculty, Gaziantep University, Gaziantep, Türkiye

[2] Department of Foreign Languages, Tourism Faculty, Higher Education Complex of Bam, Bam, Iran

Published online: 25 April 2025

1 Introduction

The emergence of ChatGPT has sparked debates regarding its pitfalls and potentials. Initially, educators and scholars perceived ChatGPT as a potential threat to traditional methods of teaching and writing (Gordon, 2023; Milmo, 2023), it has increasingly become an indispensable tool in various educational contexts. As it is clear that AI tools cannot be ignored in the near future, adapting ChatGPT to the teaching process seems to be a better option than perceiving it as a danger (Barrot, 2023; Roose, 2023). Investigating the potential uses of ChatGPT in writing instruction, can offer valuable insights and make significant contributions to practice.

Only from November 2022, when ChatGPT was launched, until six months later, 550 studies were published (Imran & Almusharraf, 2023). Some of this research (Basic et al., 2023; Guo et al., 2023; Su et al., 2023) has focused on this tool as a writing assistant. Although research on ChatGPT is burgeoning, its use in evaluating student texts remains relatively underexplored. While several studies (Baidoo-Anu & Ansah, 2023; Cotton et al., 2023; Qadir, 2022) suggest that ChatGPT can be used to evaluate student texts, there are limited experimental studies in which ChatGPT is utilized for this purpose. Among the studies no research has been conducted on the texts evaluated by ChatGPT in Turkish language. However, ChatGPT can also be used to evaluate students written texts. Barrot (2023) states that students' written assignments can be automatically graded by this AI program. Based on predetermined standards, the system gives the paper a grade and offers detailed remarks to back up that score. It is known that evaluating and assessing student texts remains one of the most challenging aspects of writing. Such that instructors also face limitations in providing quick and comprehensive feedback on student writing. Over 70% of teachers report feeling overburdened with grading and giving feedback (The Learning Agency Lab, 2023). Unfortunately, this challenge often results in reduced instructional time for teaching writing (Graham & Rijlaarsdam, 2016; Hsiang et al., 2018) due to the time-consuming nature of the task. On the other hand, students generally do not have the autonomy to determine the level and quality of their writing. Therefore, the assessment of student writing by an AI-based tool to score their texts can provide them with insight into the level of their writing and enable them to take steps to improve their skills and gain greater autonomy in their writing. To sum up it is a fact that both teachers and students benefit from ChatGPT's accurate assessment of written texts.

The goal of this study is to explore the possibility of utilizing ChatGPT as a supportive tool for both teachers and students in assessing written texts. This research holds significant implications for educational practice by refining the role of human evaluators in assessment. By leveraging ChatGPT's capabilities as a rater assistant, educators can streamline the evaluation process, focusing their expertise on higher-order aspects of student writing, such as critical thinking, creativity, and argumentation. This collaboration allows for more efficient use of educators' time and resources, enabling them to provide more targeted feedback and support to individual students. ChatGPT's accurate evaluation of students' texts will reduce the workload of teachers in the writing instruction process, allowing them to devote more time to other writing activities. Additionally, this tool could empower students to

self-assess and refine their writing skills independently. So, ensuring precise evaluation from this tool is crucial.

Accordingly, the present research is an attempt to answer the following research question:

– Is the scorer type (teachers, pre-service teachers, ChatGPT, and trained ChatGPT) a significant predictor of the likelihood that a text is categorized as having weak, moderate or advanced quality?

2 Literature review

In recent years, AI has increasingly assumed a pivotal role in the realm of writing assessment (Godwin-Jones, 2022; Gunser et al., 2022; Patout & Cordy, 2019), bringing forth a significant transformation in how educators evaluate language proficiency. Reported by numerous studies (Gupta et al., 2024; Selim, 2024), the rise of AI in educational settings has made the assessment process more accessible and efficient, enabling prompt and personalized feedback that was once time-consuming for human evaluators. With the advent of sophisticated tools like ChatGPT, AI's involvement has deepened, showing a marked evolution in capability. Initially programmed to conduct basic grammar and spell checks, these AI systems have rapidly progressed to more complex tasks. ChatGPT, for instance, has demonstrated a notable adeptness not only in developing language skills through interactive engagement but also in assessing written work, parsing through tones of semantic content and argumentative structure (Chagas, 2023; Hidayatullah, 2024; Nguyen Minh, 2024). This development signals a growing potential for AI to support both the compositional and evaluative aspects of writing, despite the drawbacks reported by Fan and Jiang (2023), presenting a dynamic shift from AI as a mere automated corrector to a meticulous contributor to language learning and assessment.

The comparative efficacy of AI versus human raters in assessing written texts has been a topic of considerable debate within academic circles (Chan, 2012; Geckïn et al., 2023; Jackaria et al., 2024). On one side, AI, exemplified by tools such as ChatGPT, offers unparalleled consistency in evaluating large volumes of text, thereby eliminating human subjectivity and fatigue (Walters, 2023). However, confirmed by Biermann et al. (2022) and Gero and Chilton (2019), AI algorithmic approach sometimes falls short in capturing the subtleties of creative expression and complex thought that characterize high-quality writing. Human raters, as argued by Geckïn et al. (2023) and Sireci and Rizavi (2000) on the other hand, offer a level of understanding and keen insight that AI has not completely replicated. While human raters excel in qualitative assessments and can appreciate the idiosyncrasies of style and voice, they are not immune to biases and inconsistencies (Sokolov, 2014). Overall, human evaluations offer a rich, contextual appreciation of text but can be variable, whereas AI provides a swift, uniform analysis but may overlook the finer points of human language.

Building upon the analysis of AI in writing assessment, the literature also explores the capabilities of pre-service teachers, when it comes to evaluating written texts. Among this group, studies indicate a developing proficiency in assessment skills (Liu, 2021; Torres, 2018), which is cultivated through educational coursework and practical experience. While these emerging educators demonstrate enthusiasm and fresh perspectives, they may lack the seasoned intuition that comes with years of hands-on teaching and grading (Dempsey et al., 2009). This can result in a degree of variance in their evaluations, raising questions about their reliability compared to more experienced educators (Street, 2003).

In contrast, in-service teachers, who have attained a higher level of practical expertise, are generally found to provide more consistent and accurate assessments. Research has shown that these veteran educators can more effectively discern subtle differences in writing quality, largely due to their extensive exposure to student work and greater familiarity with assessment rubrics (Pamela et al., 2020; Sihombing, 2016). The difference in evaluation outcomes between novice and expert raters becomes particularly evident in comparative studies. Accordingly, Nodoushan (2014), Patekar (2021), and Said et al. (2021) reported the seasoned educators' deep-rooted understanding of linguistic intricacies often leads to a more differentiated and considered approach to student writing, reflecting a level of judgment honed by years of teaching experience.

Transitioning from human raters to an examination of ChatGPT as an untrained rater unveils a different facet of assessment potential. In its default state, ChatGPT possesses the fundamental ability to evaluate written texts based on substantial training data and algorithms designed to mimic language understanding (Mahyoob et al., 2023; Sun, 2024). While this tool can rapidly process and critique vast quantities of text, delivering assessments that are free from the fatigue and partiality that can affect human raters, it is not infallible. The strengths of ChatGPT without additional training lie in its consistency and the objectivity in assessment it lends to the writing evaluation process (Algaraady & Mahyoob, 2023).

Yet, the limitations become clear when assessing advanced writing features that demand an understanding of unique stylistic elements, cultural contexts, or creative originality, i.e. areas that often pose a challenge to even the most advanced AI models. Without the benefit of specialized training or programming, untrained AI may overlook these aspects, which contribute significantly to the richness and depth of written texts (Al-Zaghir et al., 2023; Alshami et al., 2023). Thus, while ChatGPT's default capabilities indicate a strong foundation for basic assessment, the intricate layers of writing skill evaluation necessitate a subtle approach that, as yet, may be beyond the scope of AI without tailored enhancements (Alqadi et al., 2023; Luo et al., 2023).

Moving from the inherent capabilities of untrained AI, the focus shifts to how specialized training can enhance ChatGPT's effectiveness as a writing rater. When equipped with targeted training, ChatGPT demonstrates a significant improvement in its ability to assess writing, showing increased competence in identifying subtleties often missed in its default state. According to Altamimi (2023), trained AI can adapt to specific rubrics and genres of writing, providing evaluations that more closely align with the discerning viewpoint of a skilled instructor.

This training enables ChatGPT to move beyond basic error detection and into the realm of critical analysis, addressing aspects such as coherence, argument strength, and the use of discipline-specific language (Parker et al., 2023). Consequently, a trained AI not only retains the strengths of untrained models, speed and consistency, but also offers a deeper and more contextually aware assessment, narrowing the gap between machine and human evaluation of written texts.

Continuing our exploration of assessment capabilities, it is imperative to consider the role of linguistic characteristics in the Turkish language context when evaluating texts (Sönmez & Şeker, 2024). Available research, (Haerazi et al., 2018; Kaufhold, 2018) support the idea the the intricacies of any language, such as idioms, colloquialisms, and syntax, are fundamental to understanding a student's writing proficiency within that cultural setting. This understanding becomes even more crucial when dealing with a language rich in context and expression like Turkish. Therefore, the evaluator's grasp of these linguistic features significantly influences the accuracy of text assessment (Conde, 2011; Weiß et al., 2019).

While ChatGPT has demonstrated an ability to navigate complex language patterns, its proficiency in dealing with the unique aspects of the Turkish language depends largely on the quality of its training data and the scope of its programming to recognize such linguistic elements (Kesgïn et al., 2024; Kohnke et al., 2023). Human raters undoubtedly have the advantage here, as their familiarity with cultural and language-specific details inherently enables them to make more informed judgments on the quality of writing. It is through such comparative analysis (Arakawa & Yakura, 2023; Saralajew et al., 2022) between AI and human assessments that the importance of a deep understanding of language characteristics in writing evaluations, particularly for non-English texts, is accentuated.

As it was discussed earlier in the same section, the current body of research presents a comprehensive view of the evolving intersection between AI and writing assessment, highlighting the promise of tools like ChatGPT when paired with specialized training, and contrasting these with the seasoned judgments of human raters well-versed in linguistic and cultural context. Although progress has been made in enhancing AI's evaluative precision, there remains an undeniable gap in its ability to fully grasp the subtleties embedded within languages and the cultural richness that influences writing. This study seeks to bridge this gap by investigating the extent to which advanced AI can accommodate the complexities of Turkish as a specific linguistic framework and comparing its performance to that of both novice and expert human raters. Through this comparison, the study aims to illuminate areas where AI may require further refinement and where it could potentially support human assessments, contributing to a more exact understanding of AI's role in assessing writings.

3 Methodology

The following section provides a detailed description of the procedure in the data collection phase of the present study.

3.1 Selection of topics

At first, a set of twenty writing topics were selected consistent with the existing themes in Turkish language textbooks in 5th, 6th, 7th, and 8th grades in Türkiye. Subsequently, a group of twelve Turkish language teachers was requested to evaluate the appropriateness of these topics concerning middle-school students using a Likert scale ranging from 1 (completely inappropriate) to 5 (completely appropriate). These specific topics were chosen because they are aligned with common themes in middle school curricula and are familiar to students, ensuring that their focus remained on conveying information clearly and effectively rather than struggling with unfamiliar subjects. The selected topics, all of which prompted informative writing, were specifically designed to align with the assessment rubric used in this study, which focuses on the key elements of informative writing, such as clarity, organization, evidence use, and accuracy. Following the evaluation process, six topics, as shown in Table 1, that comprised the highest ratings were elected for inclusion in the study. The topics were about: "write about your favorite game" (score: 58), "write about the advantages and disadvantages of using the internet" (score: 52), "write about the ways you like animals" (score: 51), and a "write about how one can protect the environment" (score: 48). The other two topics received the same score of 46. They were about "friendship" and "an act of kindness".

3.2 Students' texts

About 165 essays were collected from 5th, 6th, 7th, and 8th grade students from four different middle schools in Türkiye. The selection of student texts followed a purposive sampling method to ensure diversity across grade levels and writing abilities. While the process was not fully randomized, care was taken to include samples representing varying proficiency levels (weak, moderate, and advanced) based on initial teacher assessments. This approach aimed to ensure that the dataset was

Table 1 Texts analysis metrics overview

QL	TN	ToT	NoW	MTS	MSoP	MSoSP	MSoLE
Weak	T1	Friendship	209	18	7	5	6
	T2	Animals	204	19	8	5	6
	T3	Game	206	18	7	5	6
Moderate	T4	Friendship	213	35	12	11	12
	T5	Internet	204	38	13	13	12
	T6	Act of Kindness	209	34	12	11	11
Advanced	T7	Nature	218	53	20	15	18
	T8	Internet	208	52	20	15	17
	T9	Friendship	206	51	20	14	17

QL Quality Level, *TN* Text Number, *ToT* Topic of Texts, *NoW* Number of Words, *MTS* Mean Total Score, *MSoP* Mean Score of Planning, *MSoSP* Mean Score of Spelling-Punctuation, *MSoLE* Mean Score of Language-Expression

reflective of the broader writing abilities within the target student population. As listed in Table 1, three levels of weak, moderate, and advanced were allocated to the texts' writing quality based on Kaldirim's (2014) grading rubric, which will be explained later. Due to the lack of high-quality texts written by the students, some revisions were made by the researchers and a Turkish language expert in some texts to assign them to the "Advanced" category. The revisions included corrections on some spelling, punctuation, grammar, planning and organization mistakes, and rewriting some parts of the texts. Finally, one of the researchers and two qualified Turkish language teachers, one holder of MA and the other one a Ph.D. in Turkish language teaching, rated totally nine texts to fall appropriately into each of the categories. The mean scores of the texts in each category are listed in Table 1.

3.3 Ranking of the students' texts

The assignment of students' texts was conducted using the analytic rubric formulated by Kaldirim (2014), available via Appendix A. The rubric included 25 items and four sub-dimensions in its original version, but was modified to 18 items and three sub-dimensions after adaptation, as explained below. The first sub-dimension, "Presentation," had seven items, including criteria related to the readability of handwriting, lettering, and page organization. However, this sub-dimension was omitted from the current study's text assignment procedure, since ChatGPT's algorithmic capabilities do not extend to the analysis of physical handwriting characteristics. The second sub-dimension, "Planning," addresses the structural organization of the text and covers seven items. The third, "Spelling-Punctuation," includes five items focusing on orthographic precision and the correct application of punctuation marks. Lastly, "Language-Expression" measures linguistic appropriateness and the clarity of the composition with six items. The adapted rubric uses a three-level scale to rate students' texts, ranging from 18 to 54. Scores from 18 to 29, 30 to 42, and 43 to 54 categorize the text as "Weak," "Moderate," and "Advanced," respectively.

3.4 Participants

In this study, four groups of raters were selected to evaluate the students' compositions based on text categories of "Weak", "Moderate", and "Advanced". Raters included experienced Turkish language teachers, pre-service teachers who were students of Turkish language teaching in Bachelor of Arts, the ChatGPT AI, and a trained version of ChatGPT. Demographic information pertaining to the teachers and pre-service teachers' group is available in Tables 2 and 3.

3.4.1 Teachers

One group of raters involved in the study included 14 Turkish Language teachers engaged in postgraduate coursework. As shown in Table 2, the raters of this group were mostly master of arts candidates, and two of them were Ph.D. candidates in the same field. The teachers were employed to evaluate the students' texts utilizing the

Table 2 Information pertaining to the teachers group

Raters	Gender	Experience	Degree	Location
T1	F	6–10 Years	Undergraduate	City
T2	F	6–10 Years	Undergraduate	Central district
T3	F	1–5 Years	Undergraduate	Village
T4	F	1–5 Years	Undergraduate	City
T5	M	6–10 Years	Undergraduate	Central district
T6	M	6–10 Years	Undergraduate	City
T7	M	16–20 Years	Undergraduate	Central district
T8	F	6–10 Years	Undergraduate	Village
T9	F	1–5 Years	Undergraduate	Village
T10	F	1–5 Years	Undergraduate	Central district
T11	F	1–5 Years	Postgraduate	City
T12	M	11–15 Years	Postgraduate	Village
T13	F	6–10 Years	Undergraduate	Central district
T14	F	< 1 Year	Undergraduate	Central district

Table 3 Information pertaining to pre-service Turkish language teachers

Raters	Gender	GPA	WCGPA
PST1	F	3.01–3.50	90 +
PST2	F	3.51 +	80–89
PST3	F	3.01–3.50	80–89
PST4	M	3.01–3.50	80–89
PST5	F	3.01–3.50	69–79
PST6	M	3.01–3.50	90 +
PST7	M	2.51–3.00	90 +
PST8	M	2.51–3.00	69–79
PST9	F	3.01–3.50	90 +
PST10	F	3.01–3.50	80–89

GPA General Point Average, *WCGPA* Writing Course General Point Average

rubric developed by Kaldirim (2014). During the rating session, to prevent potential biases in the scoring process, the teachers were presented with texts without being notified about which text is weak, moderate, or advanced in terms of text quality. The rating session spanned about 2 h. The demographic information of the teachers' group is shown in Table 2.

3.4.2 Pre-service teachers

The second group of raters comprised Pre-Service Turkish Language Teachers (see Table 3). Using the previously mentioned rubric, the students' texts were rated by 10 pre-service teachers. All of the pre-service teachers were in their final year of study. In Türkiye, the writing pedagogy course is taught in the second semester of

the third academic year; consequently, only those in the final year of the Turkish Language Teacher Training program have completed the writing pedagogy course. The completion of the writing course was considered a critical criterion for evaluating the students' texts, which is why all selected pre-service teachers were identified as being in their final year.

3.4.3 ChatGPT

The third rater in this study was ChatGPT version 3.5. ChatGPT was chosen for this study due to its advanced natural language processing capabilities, widespread accessibility, and increasing adoption in educational contexts. Compared to other AI tools, ChatGPT has demonstrated effectiveness in analyzing written texts and providing detailed feedback, making it a suitable candidate for evaluating student compositions. Data collected during the course of December 18–22, 2023. Prior to engaging ChatGPT in the evaluation process, the rating rubric was detailed within the provided prompt.

Initially, only the rubric's sub-dimensions and items were presented to ChatGPT in Turkish. For example, rubric-related prompts, translated into English, are shown below. A more complete English version is available in Appendix B.

> *I will give you a text written by a secondary school student in Türkiye. I want you to score these texts according to the following dimensions and criteria. You need to show the results in a table. The first dimension I want you to evaluate is "Planning". The second dimension is "Spelling-Punctuation" and the third dimension is "Language-Expression". In all dimensions, there are 3 levels of (1 = weak), (2 = moderate), and (3 = advanced).*
>
> *Here are the dimensions. In the first dimension, "Planning", there are 7 criteria in total. These criteria and the levels are as follows.*
> *Level 1: The title is not written.*
> *Level 2: The title is written but not suitable for the content of the subject.*
> *Level 3: A remarkable title suitable for the content of the subject has been written.*
>
> *(continued)…*

In the prompts for the untrained version of ChatGPT, no further definitions or examples were provided to guide its scoring of the texts. Following the aforementioned step, nine texts were initially presented to ChatGPT for simultaneous rating. However, due to character limitations within ChatGPT, it could not evaluate all nine texts at once. Consequently, the texts were submitted sequentially. This method was replicated ten times to enhance the reliability of ChatGPT's evaluations. To prompt ChatGPT to rate subsequent texts, the phrase "Score the following text as well" was used. At times, ChatGPT encountered confusion and uncertainty in how to proceed. For instance, it would occasionally provide feedback without a score or offer evaluations outside the defined three-level system. In such cases, the rubric and prompt were resent.

3.4.4 Trained ChatGPT

The fourth rater was a trained version of ChatGPT 3.5. To enhance ChatGPT's rating capabilities, a 51-page training document was prepared based on the previously mentioned rubric. The training dataset included a diverse range of informative texts, spanning weak, moderate, and advanced proficiency levels. Each example was annotated according to the assessment rubric, which was specifically designed for evaluating informative writing. This ensured clarity and consistency in scoring. Parameter settings, such as iteration limits, token count constraints, and response variability, were carefully configured to optimize ChatGPT's performance within this specific domain. Throughout the training process, performance was monitored through iterative test evaluations using solely informative texts, and adjustments were made to address recurring inaccuracies. Feedback loops were incorporated to ensure alignment with the rubric criteria for informative writing and improve the reliability of ChatGPT's assessment capabilities within this context. This document meticulously addressed the criteria within each sub-dimension of the rubric used for text evaluation. As illustrated in Fig. 1, specific steps were taken to train ChatGPT on the rubric's components. These steps included introducing presenting information about each item within each sub-dimension, explanation of each level of (1 = weak), (2 = moderate), and (C = advanced) for items, providing example for each item in each level, and justification of why the example has the potentiality to be put in the specified level. An example of the detailed training provided to ChatGPT follows.

As illustrated in Fig. 1, the procedure for training ChatGPT began with a prompt-based input of the rubric and ended with the presentation of texts for scoring. This process was conducted in Turkish; however, an English version of the document is available in Appendix C.

I will give you a text written by a secondary school student in Türkiye. I want you to score these texts according to the following dimensions and criteria. You need to show the results in a table. The first dimension I want you to evaluate is "Planning". The second dimension is "Spelling-Punctuation" and the third dimension is "Language-Expression". In all dimensions, there are 3 levels of (1 = weak), (2 = moderate), and (3 = advanced).

Here comes the sub-dimensions' details. First. "Planning" dimension that includes seven criteria. The criteria consist of "Title, Introduction, Main idea, Paragraph, Coherence, Transition and connections between paragraphs, and Conclusion." For each criterion, there are 3 levels of (1 = weak), (2 = moderate), and (3 = advanced). The following section provides an explanation for each criterion accompanied examples corresponding to each level.

(Continued)…

The following section presents excerpts from the prompts used to train ChatGPT.

Fig. 1 Steps taken for training chatGPT

Here comes the items in the fifth criterion. The fifth one is "Coherence." Coherence is about semantic integrity. It can be considered as the relationship between sentences and paragraphs. The characteristics of a coherent text are generally expressed as follows.

A coherent text has integrity in terms of unity between what is said at the beginning and in the end. There is continuity in theme, agents, and events. It has a central theme, and other details are included in the text in proportion to their relationship to that theme. In a coherent text, there are no unfinished thoughts or events….[1]

(Continued) …

Here comes Levels definition. For Level 1 there is no logical coherence in the text. The following is an example of level 1.

"Kişinin herhangi bir beceriyi kazanmış olduğunu ifade edebilmemiz için, söz konusu becerinin olması gereken doğruluk ölçütlerine göre yerine getirilmesi gerekir. Bununla birlikte söz konusu akıcılık olduğunda, kişinin bu düzenleme ve değiştirmeleri minimum seviyede yapması beklenir. Yazılı anlatım ürünleri olan metinler, kelime ve cümlelerin bir araya gelmesiyle oluşmaktadır…"

(Continued) …

Here is the Justification for Level 1. "In the introductory paragraph of the text given above, accuracy, fluency and writing are emphasized. In the second paragraph, …"

(Continued) …

The process of training ChatGPT on the remaining rubric details, using the prepared document, continued in the same format. Finally, texts were presented to the trained version of ChatGPT for scoring. It is worth noting that the trained ChatGPT exhibited similar confusion to its untrained counterpart. In these instances, the same corrective actions were taken as with the untrained ChatGPT.

4 Data analysis and results

The study's data was analyzed using the SPSS 26.0 software. The distribution of texts and rater groups are shown through frequencies and percentages (Table 4). Logistic regression analysis was carried out to analyze the data, given that the dependent variable was categorical. This section presents the research findings derived from data collected in line with the objectives of the present study, as detailed in Table 4.

A total of four rater groups evaluated nine texts. In the first stage, 14 teachers provided ratings for the nine texts, totally 126 ratings. Subsequently, 10 pre-service teachers rated the same set of texts, resulting in 90 ratings. ChatGPT also scored

Table 4 Findings of the groups examined in the study

Groups	Number of Texts Reviewed	%
Teacher	126	40.00
PreServiceTeacher	90	28.57
ChatGPT	90	28.57
TrainedChatGPT	9	2.86
Sum	315	100

the nine texts ten times, adding another 90 ratings. A trained version of ChatGPT rated the nine texts in the final stage. Overall, the nine texts were rated 315 times, with ratings contributed by teachers (40%), pre-service teachers and ChatGPT (each 28.57%), and trained ChatGPT (2.86%).

The present study categorized the first three texts as weak-level, the next three as moderate, and the last three as advanced to determine proficiency levels. Each group's ratings were double-checked for accurate classification. Texts aligned with their pre-determined group levels were scored as "Correct," and those that were not as "Incorrect." These classifications were then used to construct logistic regression models that gauged the precision of the group-specific predictions. The models established in relation to the research objectives are presented subsequently.

A binary logistic regression was used for the analysis due to the dichotomous nature of the dependent variable. Assumptions were carefully checked before proceeding. The first assumption concerns the balance of subjects across predictor variables. An imbalance, particularly in categorical variable combinations leading to empty cells, can result in errors and inaccurate estimates. To address this, researchers may consolidate categories, omit non-essential variables, or increase sample sizes (Field, 2005). This study encountered no empty cells requiring such actions, thereby meeting the first assumption. The second assumption involves goodness of fit tests, which may be affected if any cell has expected frequencies under five, especially if this occurs in more than 20% of cells, weakening the power of the analysis. In this case, expected values did not fall below five in more than the 20% threshold. Finally, the third assumption examines the influence of outliers on regression. Standardized residuals must be within a specific range to prevent influence on the results. Here, extreme values fell between − 0.47931 and 2.07971, which is well within the acceptable range, avoiding the critical threshold of beyond ±3 and confirming the data's appropriateness for logistic regression.

At first, for the primary model, the effect of Teachers, Pre-service Teachers, and ChatGPT rater groups on the correct classification of texts was inspected.

Considering Table 5, the correct estimation of the texts is taken as a reference point.

The chi-square value (residual chi-square) presented in Table 6 is significant ($X2 = 9.592$, $p < 0.05$, $p = 0.008$). Given the significant chi-square value, it is appropriate to proceed with the analysis (Field, 2005).

Upon examining Table 7, the significance of the model chi-square's p-value indicates a relationship between the dependent variable and the combination of independent variables. This suggests that including the independent variables in the model results in a superior prediction capability at the first stage compared to the initial model, which relied only on the constant term.

Table 5 Determination of the reference point

Correct Classification	
Original Value	Internal Value
Correct Estimate	0
Wrong Estimate	1

Table 6 Chi-Square value table for step zero

	Groups	Score	sd	*p*
Step 0	Teacher	9.592	2	0.008*
	PreServiceTeacher	3.313	1	0.069
	ChatGPT	1.268	1	0.260
X2		9.592	2	0.008*

*$p < 0.05$

Table 7 Omnibus Test on Model Coefficients

Steps	*X*2	sd	*p*
Step	9.026	2	0.011*
Block	9.026	2	0.011*
Model	9.026	2	0.011*

*$p < 0.05$

Table 8 Fit statistics of the intended model

− 2 Log likelihood	Cox & Snell R Square	Nagelkerke Square
282.239	0.029	0.047

Table 9 Coefficient estimations of the intended model

Variables	*β*	Standard Error	Wald	sd	*p*	Exp(B)	Corrected Exp (B)
Group (Teacher)			9.224	2	0.010*		
Group (PreServiceTeacher)	− 0.957	0.349	7.507	1	0.006*	0.384	2.604
Group (ChatGPT)	− 0.878	0.379	5.354	1	0.021*	0.416	2.404
Fixed	− 0.901	0.233	15.002	1	0.001*	0.406	2.463

Reviewing Table 8 reveals that the independent variable, namely the rating groups, contributes to predicting the dependent variable. Specifically, the independent variable accounts for 4.7% of the variation in correctly or incorrectly predicting the text outcome, as indicated by the Nagelkerke R Square of 0.047.

Upon examining Table 9, it was identified that the independent variables, teachers, pre-service teachers, and ChatGPT, significantly influenced the correct classification of texts ($p < 0.05$). With Teachers as the reference group, it was found that Pre-service Teachers' correct classification of texts was 2.604 times lower ($\beta = -0.957$), and ChatGPT's was 2.404 times lower ($\beta = -0.878$), as indicated by the inverse of their respective odds ratios (1/0.384 and 1/0.416). These results suggest that teachers are more accurate in text classification than both Pre-service Teachers and ChatGPT. The percentage of correct classification of the research model is given in Table 10.

When Table 10 is examined, the logistics model has shown a total of 81.7% correct classification of texts success.

Continuing the analysis, for testing the second model, the impact of Teachers, Pre-service Teachers, and Trained ChatGPT groups on the accurate classification of texts was investigated.

Upon examining Table 11, the accurate prediction of the texts is used as the reference point.

The chi-square value (residual chi-square) presented in Table 12 is not significant ($X2 = 2.648$, $p > 0.05$, $p = 0.266$). The lack of significance suggests that the analysis do need to be continued, implying that the groups do not have a significant impact on the classification of the texts (Field, 2005).

Ultimately, Teachers and Pre-service Teachers were categorized as the "Human" rater group, while ChatGPT and Trained ChatGPT were placed in the "Machine" rater group, aiming to discern the effect of these collective groups on the correct classification of texts.

When looking at Table 13, the correct prediction of texts is utilized as the benchmark for reference.

The chi-square value (residual chi-square) presented in Table 14 is significant ($X2 = 10.582$, $p < 0.05$, $p = 0.001$). The significance of this chi-square value justifies the

Table 10 Percentage of post-model classification

		Predicted		
		guess		Percentage Correct
		True	Mistake	
Actual Situation	True	250	0	100.0
	Mistake	56	0	0.0
Success Percentage				81.7

Table 11 Determination of the reference point

Correct Classification	
Original Value	Internal Value
Correct Estimate	0
Wrong Estimate	1

Table 12 Chi-square value table for step zero

	Variables	Score	sd	*p*
Step 0	Teacher	2.648	2	0.266
	PreServiceTeacher	0.316	1	0.574
	TrainedChatGPT	0.006	1	0.939
X2		2.648	2	0.266

*$p < 0.05$

Table 13 Determination of the reference point

Correct classification	
Original Value	Internal Value
Correct Estimate	0
Wrong Estimate	1

Table 14 Chi-square value table for step zero

	Variables	Score	sd	*p*
Step 0	Group	10.582	1	0.001*
*X*2		10.582	1	0.001*
*$p < 0.05$				

Table 15 Omnibus test on model coefficients

Steps	*X*2	sd	p
Step	10.029	1	0.002*
Block	10.029	1	0.002*
Model	10.029	1	0.002*
*$p < 0.05$			

Table 16 Fit statistics values of the intended model

− 2 Log likelihood	Cox & Snell R Square	Nagelkerke Square
293.812	0.031	0.051

continuation of the analysis, indicating that the rater groups potentialy have a measurable effect on the classification of the texts.

Upon reviewing Table 15, the significance of the *p* value for the model chi-square indicates that there is a relationship between the dependent variable and the collective independent variables. This implies that incorporating the independent variables into the model enhances its predictive accuracy compared to an initial model that relies solely on the constant term.

After examining Table 16, it was determined that the independent variable (i.e., groups) in the model is the explanatory factor for predicting the dependent variable. Specifically, the independent variables included in the model account for 5.1% of the variance in correctly or incorrectly classifying the text, as indicated by the Nagelkerke R Square value of 0.051.

Upon review of Table 17, it was discovered that the independent variables, human and machine, had a significant effect on the correct classification of texts ($p < 0.05$). With the human category set as the reference group, the research findings indicated that humans' correct classification rating of the texts ($\beta = 0.943$) were 2.569 times higher than those of machines. This result suggests that humans outperform

Table 17 Coefficient estimations of the intended model

Variables	β	Standard Error	Forest	sd	p	Exp(B)
Group (Human)	0.943	0.286	10.173	1	0.001*	2.569
Constant	− 1.825	0.197	85.999	1	0.001*	0.161

machines in correctly classifying the texts. The percentage of correct classification of the research model is given in Table 18.

On inspecting Table 18, it is evident that the logistic model has achieved a total correct classification accuracy of 81.3%.

These results demonstrate the significant influence of human raters on text categorization accuracy ($\beta = 0.943$, $p < 0.001$). Specifically, a one-unit increase in the human rater group variable is associated with a 2.569 times higher likelihood of correct categorization compared to machines. This suggests that human evaluators excel at capturing complex elements of text that AI systems might overlook. Similarly, pre-service teachers demonstrated lower performance ($\beta = -0.957$, $p = 0.006$) with a 2.604 times lower likelihood of correct classification compared to teachers. ChatGPT also displayed reduced accuracy ($\beta = -0.878$, $p = 0.021$) with a 2.404 times lower likelihood of correct classification compared to teachers.

While ChatGPT performed well on certain objective measures, as reflected in its ability to handle surface-level text features, the human evaluators consistently outperformed ChatGPT on measures related to subtle understanding and contextual interpretation, as evidenced by their significantly higher coefficients and odds ratios. These findings highlight the complementary strengths of AI-driven tools and human evaluators in text assessment.

The findings of the data analysis reveal distinct differences in text classification accuracy among various evaluator groups. The primary model highlighted that teachers outperform both pre-service teachers and ChatGPT, with a verifiable impact on correct text classification as evidenced by significant chi-square values. Teachers demonstrated a strong ability to capture complex elements of text organization and linguistic expression, contributing to their higher classification accuracy. Pre-service teachers, while showing promise, displayed variability in applying the rubric consistently, likely due to their limited practical experience. ChatGPT,

Table 18 Post-model classification percentage

		Predicted		
		guess		Percentage Correct
		True	Mistake	
Actual Situation	True	250	0	100.0
	Mistake	59	0	0.0
Success Percentage				81.3

despite its efficiency in identifying structural and surface-level errors, struggled with interpreting contextual subtleties and stylistic complexities, which are often better assessed by human evaluators.

In contrast, the second model showed no significant differences among the teachers, pre-service teachers, and trained ChatGPT, indicating that training may not notably enhance ChatGPT's ability to classify texts when compared to human evaluators. The final model offered a stark contrast between groups labeled 'Human' and 'Machine,' with humans exhibiting significantly higher accuracy in text classification than machines. These results, with a logistic model accuracy rate of over 81%, underscore the complexity of text classification tasks and the current limitations of even trained AI systems in matching human performance. The interpretative strength of humans in subtle tasks such as text classification remains superior, reaffirming the indispensable role of human judgment in scoring texts.

The findings suggest that while ChatGPT offers consistency and efficiency in text assessment, human evaluators, particularly experienced teachers, demonstrate a stronger capacity for capturing complex elements of writing, such as argument strength and stylistic clarity. These strengths translate into a significantly higher likelihood of correct text classification. Pre-service teachers, though promising, require more practical experience to match the accuracy of experienced educators. These results highlight the potential for an integrated approach where AI handles initial evaluations, and human raters provide deeper qualitative insights, ensuring both efficiency and accuracy in assessment practices.

5 Discussion

This study's findings show that human raters are better at accurately judging written texts than AI systems like ChatGPT. Although ChatGPT, especially when it has been trained, is impressive, it still does not match the unique skills of human teachers and those training to become teachers. These individuals have a better track record than AI when it comes to the detailed and sometimes subtle work of telling texts apart. This point supports what other recent studies have also reported (Wang & Demszky, 2023; Yang et al., 2023). They concluded that human judgment retains its critical role in discerning the intricacies of textual content. These results back up the idea that even with the progress made in AI, understanding the fine points of language and the setting it is used in is still something humans are better at. The numbers are clear: humans are much better at correctly figuring out what category a text falls under, with around 2.6 times the accuracy of AI algorithms used for the same task. In the context of our study, the numbers indicate that human raters, both teachers and pre-service teachers, demonstrated approximately 2.6 times higher accuracy in categorizing

texts compared to AI algorithms (ChatGPT and trained ChatGPT). This finding highlights the current limitations of AI in handling complex aspects of text evaluation within the specific rubric and dataset used in our research. This not only shows where AI is at right now but also echoes what has found in recent research (Son et al., 2023; Tseng & Warschauer, 2023), that when it comes to catching the details of written material, human insight is still essential.

The clear difference in how well human raters and AI systems like ChatGPT rate texts reveals a lot about the use of AI in assessing writing. The study shows us that even with fast-growing technology, there is something about the way humans understand the fine points of education that AI has not mastered. AI has the potential to tailor education to each learner, as described by researchers like Ngo (2023) and Rahman and Watanobe (2023), but this study makes it clear that AI on its own might not be ready to fully take on the detailed, personal parts of language that humans can rate so well.

The standout performance of humans over AI in categorizing texts strongly supports using both together for the best educational outcomes. AI could be game-changing for language learning, as it can create tailored assissting paths that suit each teacher's needs as a rater, as Lutskovskaia et al. (2019) have noted. But we should not forget how good teachers are at picking up on the delicate language details and cultural subtlities. This study's results are showing just how crucial humans are in rating students' writing, with a deeper ability to interpret that AI has not reached yet. Even though AI has shown it can rate and assess written content, the fact that human raters are so much better at precisely sorting texts means that their role at the heart of grading should stay strong (Mageira et al., 2022).

Moreover, this study adds important perspectives to the ongoing conversation about how to use AI responsibly and effectively in assessing written texts. As AI tools become more common, and as schools use more data to make decisions, it is really important for teachers and school leaders to make sure that AI does not unintentionally cause further bias or unfairness, something researchers like Ntoutsi et al. (2020) and Paravattil and Wilby (2019) have warned about. So, even though AI has a lot to offer in helping education reach farther and work better, the findings from this study remind us that it is important for humans to stay in charge. This is to make sure that fairness, quality, and honesty stay at the heart of how written texts are rated, a concern that have been shared by many experts (Lu, 2019; Mondal, 2019). The natural instincts and decision-making skills that humans bring to the table are still essential, making sure that as AI becomes a bigger part of education, it enhances how we learn rather than takes away from the crucial human skill involved in teaching and evaluating.

When we look at the present study findings alongside other research, we see a clear trend in how AI fits into education: technology is helpful, but it does not replace the unique skills human bring to the table (Foster, 2019; Wang & Davier, 2014). Our findings, showing that human still do better than AI in figuring out

what category a text belongs to, are similar to what others like Lutskovskaia et al. (2019) and Selim (2024) have found. They see the benefit in AI tools that can adapt educational content to match a student's level. But these tools do not yet fully understand and judge the way humans do, which is a complex task. This gap tells us that while AI is great for assessing written texts, humans are the ones we cannot do without when it comes to evaluating and interpreting the quality of their work. This supports the cautious approach the education field is taking towards AI adoption, which researchers like Luan et al. (2020) and Mafara and Abdullahi (2024) have also noticed.

Present study shows that humans significantly outperform machines in text classification accuracy which is in contrast with more optimistic perspectives presented by Zhang et al. (2024), Wang et al. (2023), and Tseng and Warschauer (2023) reporting AI as a transformative force in scoring tasks and texts. While above-mentioned studies speak to AI's potential to revolutionize language education through increased accessibility and judging students' performance in a productive skill such as writing, the current study indicates that this revolution has not yet usurped the need for traditional human-led educational strategies. Indeed, it becomes apparent that the synchronous growth of AI cannot disregard the existing infrastructural framework of human educators, whose cognitive and empathic strengths continue to steer the course of effective teaching and learning. This comparative analysis illuminates an educational landscape where, despite AI's prominent strides, the empathetic and intuitive qualities of human instruction and assessment remain at its core, tempering the notion that AI might soon become autonomous in domains that demand deep interpretive and evaluative abilities.

The implications of the study's findings for educational practices are significant, presenting a refined perspective on the interplay between human evaluators and AI in the realm of education. The persisting edge that human evaluators hold in the domain of text classification accuracy charters a clear instruction for educational institutions to carefully calibrate the integration of AI within their pedagogical strategies. Despite AI's advancements, it is pivotal to recognize, as highlighted in frequent studies (Japoshvili-Ghvinashvili & Suleman, 2023; Yu, 2023), that it is not a substitute for human expertise but rather a complementary tool that can bolster the effectiveness of educational interventions. AI's potential in automating administrative tasks and offering personalized support, as highlighted by Karakose (2023), is invaluable, and yet human teachers remain the mainstay for providing essential support in assessing written texts.

In the classroom, AI can be a helpful assistant to teachers, taking care of routine grading and highlighting which students might need extra help. This lets teachers give their attention to the things that really need a personal touch, like sparking lively class debates, encouraging students to think deeply, and looking after the emotional and social growth of every kid in the class. By teaming up AI's strengths with the essential skills of teachers, we could make school better for students and give teachers a break from paperwork. This extra time could

then go into creating new teaching methods and getting to know students better. This approach aligns with the propositions set forth by Eaton (2017), who advocates for an incorporation of AI's data-driven insights with the teacher's interpersonal skills, and Goel and Joyner (2017), who emphasize the importance of teacher training in AI literacy to ensure educators are well-equipped to implement AI tools within their teaching repertoire. By using these strategies, teachers can guide how AI helps with evaluating written work, making sure to keep and improve the unique personal touches that only a human teacher can provide when it comes to understanding and giving feedback on student writing. Therefore, the Hybrid AI-Human Writing Assessment Model (HAHWAM), illustrated in Fig. 2, is proposed to optimally integrate ChatGPT as a rater assistant.

Figure 2 presents a proposed framework, the Hybrid AI-Human Writing Assessment Model, which envisions a collaborative approach to text evaluation. In this model, AI tools efficiently handle initial evaluations, focusing on objective criteria, while human raters subsequently provide deeper qualitative insights. Future research could explore the practical implementation and refinement of this HAHWAM framework, investigating its impact on student writing outcomes and assessing its scalability across diverse educational contexts and languages. Further research directions include optimizing workflows for human-AI collaboration within this hybrid model, examining the long-term effects of AI-influenced evaluation on student writing proficiency, and exploring the potential of such integrated assessment models to enhance both accuracy and efficiency in large-scale assessment scenarios.

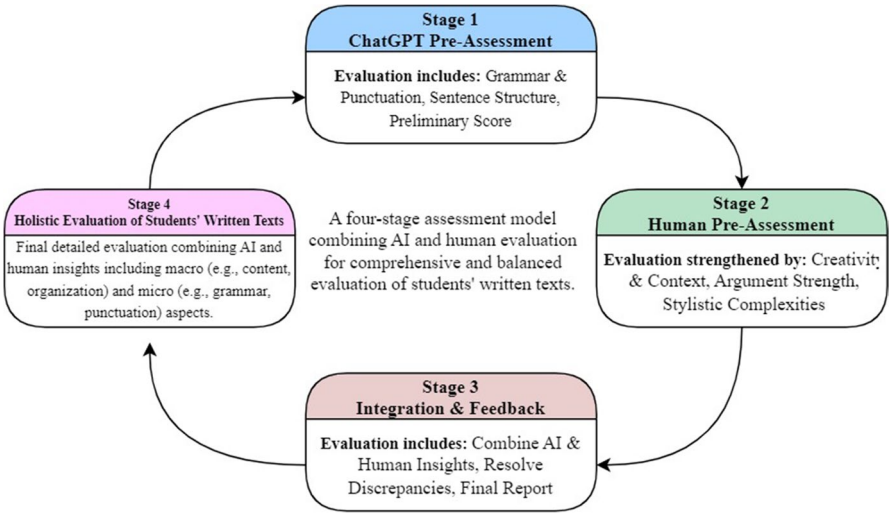

Fig. 2 Hybrid AI-Human Writing Assessment Model (HAHWAM). Note: Stages of *HAHWAM* in evaluation of student's writing. This model illustrates the progression from AI-based preliminary evaluation to human evaluation, integration of insights, and final holistic

5.1 Limitations

The design and methodology of this research provide a robust framework for evaluating the efficacy of human evaluators versus AI in assessing writings, a strength that is particularly notable given the complexity of the task. By employing a comparative analysis that involved both ChatGPT and human evaluators, the study has outlined the capabilities and limitations of AI in educational settings. A rigorous approach, designed to minimize variables, has offered clear insights into the refined domain of writing assessment. However, it is important to acknowledge certain limitations that might have influenced the outcomes. The scope of language diversity in the text used for classification, possible bias in the selection of evaluators, and the limited iterations of AI training could affect the generalizability of the results. The evolving nature of AI algorithms also means that ongoing advancements may alter the landscape, rendering these findings a snapshot within a rapidly advancing field. Moreover, the study relies on the current state of AI technology, which is in constant flux, voicing the necessity for continued research and evaluation to keep abreast of technological progress and its implications for educational practices.

6 Conclusion

This study shines a light on a key moment in how we evaluate students' texts: it shows that while AI like ChatGPT is really advanced, it is still not as good at grading texts as human teachers are. Human teachers seem to understand and think about the context of written work in a way that AI has not managed to do yet. The results emphasize that human insight remains indispensable in assessing writing. A balanced integration, where AI tools complement human expertise, appears to be the most effective approach, leveraging the strengths of both for optimal outcomes. As we look ahead, we will need to be smart and practical about how we use AI and humans together in the classroom.

This study points to a critical discussion about the role of AI in education. It suggests that technology should be used to boost the essential human aspects of assessment, not replace them. Finding the right balance is key when bringing AI into the assessing writing, making sure not to undervalue teacher expertise, especially when grading and giving feedback on written texts. A practical recommendation would be to establish a hybrid assessment framework where ChatGPT is used for initial evaluation, focusing on objective criteria such as grammar, punctuation, and structural consistency, while human evaluators provide deeper qualitative insights into text organization, creativity, and contextual appropriateness. Training programs for educators should also include modules on integrating AI tools effectively into assessment workflows to maximize their potential.

Appendix A: An English Version of the Rubric Developed by Kaldırım(2014)

Table 19 An english version of the rubric developed by Kaldirim (2014)

	1	2	3
Planning	The title has not been written	The title has been written but not appropriate to the content of the topic	An attention-grabbing title appropriate to the content of the topic is written
	The text does not begin with an appropriate introduction to the topic	The text lacks a clear overview of the topics to be covered, which can make it difficult for readers to understand its purpose	The text begins with an engaging introduction that accurately previews the content to follow
	The text lacks a clear central idea	The unity of the paragraphs is deficient	The text is organized around a main idea and is supported by relevant supporting ideas
	The single paragraph meanders between unrelated ideas, lacking a clear focus. There is no logical coherence in the text	The unity of the paragraphs is deficient	Paragraphs maintain a complete unity of subject matter within themselves
	There is no logical coherence in the text	The topic was approached with logical consistency, but some errors were made that did not impact the overall understanding of the text	The subject is handled with logical consistency
	The text lacks transitions, making it difficult to see the connection between paragraphs	There are deficiencies in transitions and connections between paragraphs	Transitions and connections between paragraphs are used effectively
	The concluding section does not effectively summarize the issue or provide closure	The text has a conclusion, but it is not effective	There is a conclusion section that appropriately concludes the text and impresses the reader
Spelling-punctuation	Numerous spelling mistakes are present throughout the text. (10 + mistakes)	Some spelling mistakes were made. (5–9 mistakes)	Most of the words are spelled correctly. (0–4 mistakes)
	The text contains multiple capitalization errors. (10 + mistakes)	Some capitalization mistakes were made. (5–9 mistakes)	Capitalization is correct in many places in the text. (0–4 mistakes)
	There are multiple instances of incorrect preposition and conjunction choices throughout the text. (5 + mistakes)	Some mistakes were made in the spelling of prepositions and conjunctions. (3–5 mistakes)	Most prepositions and conjunctions are written correctly. (0–2 mistakes)
	The text contains numerous punctuation errors. (10 + mistakes)	Some punctuation errors were made. (5–9 mistakes)	Most of the punctuation marks are used accurately. (0–4 mistakes)

Table 19 (continued)

	1	2	3
	The text's meaning is hampered by frequent grammatical errors. (5 + mistakes)	Some grammar mistakes were made. (3–5 mistakes)	Most grammar rules are used correctly, aiding in the understanding of the text. (0–2 mistakes)
Language-expression	The author's unclear writing makes it difficult to understand the intended meaning	The author has expressed their thoughts, but some points remain unclear in the text	The author has clearly expressed his thoughts
	The text lacks supporting evidence such as quotations, examples, or analogies	Quotations, examples, and analogies expressing the situation or event are used incompletely or incorrectly	Quotations, examples, and analogies expressing the situation or event are used effectively and appropriately for the content
	Several words are used incorrectly, impacting the clarity and accuracy of the text. Sentences are not diversified	The word choice is correct, but there is no variety	Words are skillfully chosen and used in the correct sense
	Sentences are not diversified	Sentences are not diversified enough	Sentences make the test effective and varied
	The text contains unnecessary repetitions	There is a little unnecessary repetition in the text	The text is written without repetition
	There is no subject-predicate unity in sentences	There are some sentences in the text where there is no subject-predicate unity	All sentences have subject-predicate unity

Appendix B: An Example of a Prompt Given to ChatGPT

I will give you a text written by a secondary school student in Türkiye. I want you to score these texts according to the following dimensions and criteria. You need to show the results in a table. The first dimension I want you to evaluate is "Planning". The second dimension is "Spelling-Punctuation" and the third dimension is "Language-Expression". In all dimensions, there are 3 levels of (1 = weak), (2 = moderate), and (3 = advanced).

Here are the dimensions. In the first dimension, "Planning", there are 7 criteria in total. These criteria and the levels are as follows.

The first criterion and levels are as follows
Level 1: The title is not written.
Level 2: The title is written but not suitable for the content of the subject.
Level 3: A remarkable title suitable for the content of the subject has been written.
…

The seventh criterion and levels are as follows
Level 1: The text does not have a section that concludes the topic with appropriate expressions.
Level 2: There is a conclusion section in the text. But the writing is not effective.
Level 3: There is a conclusion section that concludes the writing with appropriate expressions and affects the reader.

In the second dimension, "Spelling-Punctuation", there are 5 criteria in total. These criteria and the levels are as follows.

The first criterion and levels are as follows:
Level 1: Many mistakes were made in the spelling of words (10+ mistakes)
Level 2: Some mistakes were made in the spelling of words (5-9 mistakes)
Level 3: Most of the words are spelled correctly (0-4 errors)
…
The fifth criterion and levels are as follows:
Level 1: Many mistakes were made in the use of grammar rules that negatively affected the meaning (5+ mistakes)
Level 2: Some mistakes were made in the use of grammar rules (3-5 mistakes)
Level 3: Most of the grammar rules are used correctly and help to understand the text correctly (0-2 errors)

In the third dimension, "Language-Expression", there are 6 criteria in total. These criteria and the levels are as follows.

The first criterion and levels are as follows:
Level 1: The author has not expressed his/her thoughts clearly. The reader cannot understand what is intended to be conveyed.
Level 2: The author has expressed his/her thoughts, but there are some points that remain unclear in the text.
Level 3: The author has clearly expressed his/her thoughts.
…
The sixth criterion and levels are as follows:
Level 1: There is no subject-predicate unity in the sentences.
Level 2: There are some sentences in the text that do not have subject-predicate unity.
Level 3: Sentences have subject-predicate unity.
Here is the text that I want you to evaluate according to the above criteria. Make the evaluation separately for each dimension and show it in a table.

"The text to be scored was given here"

Appendix C: An Example of a Prompt Given to Trained ChatGPT

I will give you a text written by a secondary school student in Türkiye. I want you to score these texts according to the following dimensions and criteria. You need to show the results in a table. The first dimension I want you to evaluate is "Planning". The second dimension is "Spelling-Punctuation" and the third dimension is "Language-Expression". In all dimensions, there are 3 levels of (1 = weak), (2 = moderate), and (3 = advanced).

When evaluating a written text there are 3 levels of (1 = weak), (2 = moderate), and (3 = advanced) for the "Coherence" criterion. In the following section, the "Coherence" criterion is explained, and examples corresponding to the (1 = weak), (2 = moderate), and (3 = advanced) levels of this criterion are presented.

Coherence is about semantic integrity. It can be considered as the relationship between sentences and paragraphs. The characteristics of a coherent text are generally expressed as follows. A coherent text has integrity in terms of unity between what is said at the beginning and in the end. There is continuity in theme, agents, and events. It has a central theme, and other details are included in the text in proportion to their relationship to that theme. In a coherent text, there are no unfinished thoughts or events. A coherent text revolves around a central topic, with all details meticulously arranged based on their relevance to this theme. It leaves no room for ambiguity, ensuring a logical flow of ideas and a seamless connection between emotions, thoughts, and events. Contradictions are nonexistent, as each new piece of information builds upon the previous one, contributing to a unified whole. Repetition is avoided, with every element serving a distinct purpose. The absence of any unit would create a noticeable gap, as each is integral to the text's overall meaning. A coherent text anticipates the reader's need for clarity, providing all necessary information to bridge any potential comprehension gaps. Finally, it maintains a consistent style and tone throughout, harmonizing form and content.

Level 1: There is no logical coherence in the text.

The Example of Level 1: "Kişinin herhangi bir beceriyi kazanmış olduğunu ifade edebilmemiz için, söz konusu becerinin olması gereken doğruluk ölçütlerine göre yerine getirilmesi gerekir. Bununla birlikte söz konusu akıcılık olduğunda, kişinin bu düzenleme ve değiştirmeleri minimum seviyede yapması beklenir. Yazılı anlatım ürünleri olan metinler, kelime ve cümlelerin bir araya gelmesiyle oluşmaktadır.

Yazma işlekliği kavramının iki boyutu olduğu anlaşılmaktadır. Kelimeler olmadan düşüncelerin aktarılması veya tam anlamıyla ifade edilmesi olanaksızdır. Kişi zihninde paylaşmaya hazır bir söz oluşturduktan sonra onu karşıdakilere iletmek üzere dilin yazılı ya da sözlü boyutunu kullanır. Cümle "Bir yargıyı bildirmek için tek başına çekimli bir fiil veya çekimli bir fiille kullanılan kelimeler dizisi, tümce." şeklinde tanımlanmaktadır. Yazma ve konuşma, anlatma ihtiyacımızı karşılayan dil becerileri olmaları bakımından bazı ortak özelliklere sahip olsalar da birbirinden ayrıldıkları pek çok nokta vardır.

Kısaca yazılı dile taşınan düşüncelerin rastgele bir biçimde sıralanmış, anlam birliği oluşturmaktan uzak, birbirinden kopuk, düzensiz olmaları yazılı metin oluşturmanın önündeki en büyük engeldir."

The Explanation of the Example of Level 1: In the introductory paragraph of the text given above, accuracy, fluency and writing are emphasized. The second paragraph, intended as the development section, posits that writing processes, sentence definitions, and writing/speaking skills share certain commonalities and distinctions. However, the subsequent conclusion jumps to the assertion that the primary challenge in writing lies in the irregular expression of thoughts. This leap highlights a lack of coherence: the ideas presented lack a logical connection and thematic unity. A coherent text requires a clear progression of ideas, with smooth transitions between sentences and paragraphs. In contrast, this text introduces a new concept with almost every sentence, disrupting the flow and undermining its logical coherence. Therefore, it falls short of the coherence standards expected in the Plan and Organization dimension, meriting a level 1 evaluation.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This research received no external funding.

Data availability Data will be available upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest relating to the research, authorship, and/or publication of this study.

References

Al-Garaady, J., & Mahyoob, M. (2023). *ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. Arab World English Journals, 9*(CALL). https://ssrn.com/abstract=4519092

Al-Qadi, R., Al-Rbaiyan, A., Al-Rumayyan, N., Al-Qahtani, N., & Najjar, A. B. (2023). *Exploring the user experience and the role of ChatGPT in the academic writing process. 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 1082–1089. https://doi.org/10.1109/CSCE60160.2023.00180

Al-Shami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., & Zayed, T. (2023). Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems, 11*(7), 351. https://doi.org/10.3390/systems11070351

Al-Tamimi, A. B. (2023). Effectiveness of ChatGPT in essay autograding. *2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, 102–106. https://doi.org/10.1109/iCCECE59400.2023.1023854

Al-Zaghir, Z., Al-Naqbi, N. M., Matroud, A. A., & Abdalgader, K. (2023). Exploring opportunities and challenges of using ChatGPT in professional writing instruction. *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 1–6. https://doi.org/10.1109/TALE56641.2023.10398234

Arakawa, R., & Yakura, H. (2023). *AI for human assessment: What do professional assessors need?* In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)* (Article 378, pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3544549.3573849

Baidoo-Anu, D., & Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*. https://doi.org/10.61969/jai.1337500

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Basic, Z., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications, 10.* https://doi.org/10.1057/s41599-023-02269-7

Biermann, C., Ma, N. F., & Yoon, D. (2022). From tool to companion: Storywriters want AI writers to respect their personal values and writing strategies. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)* (pp. 1209–1227). Association for Computing Machinery. https://doi.org/10.1145/3532106.3533506

Chagas, L. (2023). ChatGPT in foreign language learning: A short experiment with tourism students. *International Research in Higher Education, 8*(2). https://doi.org/10.5430/irhe.v8n2p39

Chan, K. Y. (2012). *A comparison of automated scoring engines and human raters on the assessment of English essay writing* (Doctoral dissertation, James Cook University). https://doi.org/10.25903/n4vp-4087

Conde, T. (2011). Translation evaluation on the surface of texts: A preliminary analysis. *The Journal of Specialised Translation, 15*, 69–86.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing, 14*, 38–61. https://doi.org/10.1016/j.asw.2008.12.003

Eaton, E. (2017). Teaching integrated AI through interdisciplinary project-driven courses. *AI Magazine, 38*, 13–21. https://doi.org/10.1609/aimag.v38i2.2730

Fan, Y., & Jiang, F. (2023). Uncovering the potential of CHATGPT for discourse analysis in dialogue: An empirical study. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.08391

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). Sage Publications.

Foster, E. (2019). Study examines teachers' perceptions of student achievement data. *The Learning Professional, 40*(3), 20–23.

Geckin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. *Journal of Educational Technology and Online Learning, 6*(4), 1096–1108. https://doi.org/10.31681/jetol.1336599

Gero, K., & Chilton, L. B. (2019). *Metaphoria: An algorithmic companion for metaphor creation.* In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (Paper 296, pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300526

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology, 26*(2), 5–24.

Goel, A. K., & Joyner, D. A. (2017). Using AI to teach AI: Lessons from an online AI class. *AI Magazine, 38*(2), 48–59. https://doi.org/10.1609/aimag.v38i2.2732

Gordon, C. (2023). *How are educators reacting to ChatGPT?* Retrieved January 3, 2024, from www.forbes.com

Graham, S., & Rijlaarsdam, G. (2016). Writing education around the globe: Introduction and call for a new global analysis. *Reading and Writing, 29*. https://doi.org/10.1007/s11145-016-9640-1

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., Çakir, D. C., & Gerjets, P. (2022). *The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors?* In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)* (pp. 60–61). Association for Computational Linguistics.

Guo, K., Li, Y., Li, Y., & Chu, S. (2023). Understanding EFL students' chatbot-assisted argumentative writing: An activity theory perspective. *Education and Information Technologies, 29*. https://doi.org/10.1007/s10639-023-12230-5

Gupta, S., Dharamshi, R. R., & Kakde, V. U. (2024). *An impactful and revolutionized educational ecosystem using generative AI to assist and assess the teaching and learning benefits, fostering the post-pandemic requirements. 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)* (pp. 1–4). https://doi.org/10.55248/gengpi.5.0724.1734

Haerazi, H., Irwansyah, D., Juanda, J., & Azis, Y. A. (2018). Incorporating intercultural competences in developing English materials for writing classes. *Journal of Language Teaching and Research, 9*, 540–547. https://doi.org/10.17507/jltr.0903.13

Hidayatullah, E. (2024). Exploring interactivity and engagement: Improving writing skills with ChatGPT for fun learning. *Journal of Language, Literature, and Teaching, 5*(3). https://doi.org/10.35529/jllte.v5i3.16-26

Hsiang, T. P., Graham, S., & Wong, P. M. (2018). Teaching writing in grades 7–9 in urban schools in Chinese societies in Asia. *Reading Research Quarterly, 53*(4), 473–507. http://www.jstor.org/stable/26622531.

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*, 1–14. https://doi.org/10.30935/cedtech/13605

Jackaria, P. M., Hajan, B. H., & Mastul, A.-R. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research, 23*(2). https://doi.org/10.26803/ijlter.23.2.23

Japoshvili-Ghvinashvili, M., & Suleman, D. N. (2023). Assisting ELT teachers: Designing activities for the use of ChatGPT in teaching and learning. *Pakistan Journal of Multidisciplinary Innovation, 2*(1). https://doi.org/10.59075/pjmi.v2i1.219

Kaldirim, A. (2014). *The effect of 6+1 analytical writing and evaluation model on the written expression skills of sixth-grade secondary school students* (Master's thesis, Dumlupınar University, Institute of Educational Sciences, Department of Turkish Education).

Karakose, T. (2023). The utility of ChatGPT in educational research—Potential opportunities and pitfalls. *Educational Process International Journal, 12*(2). https://doi.org/10.22521/edupij.2023.122.1

Kaufhold, K. (2018). Creating translanguaging spaces in students' academic writing practices. *Linguistics and Education, 45*(3). https://doi.org/10.1016/j.linged.2018.02.001

Kesgïn, H. T., Yuce, M. K., Dogan, E., Uzun, M. E., Uz, A., Seyrek, H. E., Zeer, A., & Amasyali, M. F. (2024). Introducing cosmosGPT: Monolingual training for Turkish language models. arXiv:2404.17336. https://doi.org/10.48550/arXiv.2404.17336

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(3), 537–550. https://doi.org/10.1177/00336882231162868

Liu, L. (2021). Scoring judgment of pre-service EFL teachers: Does writing proficiency play a role? *The Asia-Pacific Education Researcher, 31*(1), 333–343. https://doi.org/10.1007/s40299-021-00575-9

Lu, X. (2019). An empirical study on the artificial intelligence writing evaluation system in China CET. *Big Data, 7*(2), 121–129. https://doi.org/10.1089/big.2018.0151

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.-C. (2020). Challenges and future directions of big data and artificial intelligence in education [Review]. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.580820

Luo, Z., Xie, Q., & Ananiadou, S. (2023). ChatGPT as a factual inconsistency evaluator for abstractive text summarization. arXiv:2303.15621. https://doi.org/10.48550/arXiv.2303.15621

Lutskovskaia, L. Y., Shoustikova, T., & Udina, N. (2019). AI-based tools for foreign language training: Opinions from different audiences.

Mafara, R. M., & Abdullahi, S., Suleiman. (2024). Adopting artificial intelligence (AI) in education: Challenges & possibilities. *Asian Journal of Advanced Research and Reports, 18*(2). https://doi.org/10.9734/ajarr/2024/v18i2608

Mageira, K., Pittou, D., Papasalouros, A., Kotis, K. I., Zangogianni, P., & Daradoumis, A. (2022). Educational AI chatbots for content and language integrated learning. *Applied Sciences, 12*(2). https://doi.org/10.3390/app12073239

Mahyoob, M., Algaraady, J., & Alblwi, A. (2023). Proposed framework for human-like language processing of ChatGPT in academic writing. *International Journal of Emerging Technologies in Learning, 18*(14), 282–293. https://doi.org/10.3991/ijet.v18i14.41725

Milmo, D. (2023, April 9). Italy's privacy watchdog bans ChatGPT over data breach concerns. *The Guardian*. https://www.theguardian.com/technology/2023/mar/31/italy-privacy-watchdog-bans-chatgpt-over-data-breach-concerns

Mondal, K. (2019). A synergy of artificial intelligence and education in the 21st-century classrooms. *International Conference on Digitization (ICD), 2019*, 68–70.

Ngo, T. C. T. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning (iJET), 17*(8). https://doi.org/10.3991/ijet.v18i17.39019

Nguyen Minh, A. (2024). Leveraging ChatGPT for enhancing English writing skills and critical thinking in university freshmen. *Journal of Knowledge Learning and Science Technology, 3*(2), 51–62. https://doi.org/10.60087/jklst.vol3.n2.p62

Nodoushan, M. (2014). Assessing writing: A review of the main trends. *Studies in English Language and Education, 1*(2), 116–125. https://doi.org/10.24815/siele.v1i2.1831

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wires Data Mining and Knowledge Discovery, 10*, e1356. https://doi.org/10.1002/widm.1356

Pamela, K., Refnaldi, & Zaim, M. (2020). Teachers' needs for authentic assessment to assess writing skill at grade X of senior high schools in Tanah Datar. *Proceedings of the Eighth International Conference on Languages and Arts (ICLA-2019)*. https://doi.org/10.2991/assehr.k.200819.005

Paravattil, B., & Wilby, K. J. (2019). Optimizing assessors' mental workload in rater-based assessment: A critical narrative review. *Perspectives on Medical Education, 8*, 339–345. https://doi.org/10.1007/s40037-019-00535-6

Parker, J., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *The Journal of Nursing Education, 62*(12), 721–727. https://doi.org/10.3928/01484834-20231006-02

Patekar, J. (2021). A look into the practices and challenges of assessing young EFL learners' writing in Croatia. *Language Testing, 38*, 456–479. https://doi.org/10.1177/0265532221990657

Patout, P.-A., & Cordy, M. (2019). Towards context-aware automated writing evaluation systems. *Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence*.

Qadir, J. (2022). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *TechRxiv*. https://doi.org/10.36227/techrxiv.21789434

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9). https://doi.org/10.3390/app13095783

Roose, K. (2023). Don't ban ChatGPT in schools. Teach with it. *The New York Times*. Retrieved January 3 from https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

Said, R. R., Jusoh, Z., Md. Sabil, A., & Othman, S. (2021). Teacher readiness in assessing students for Malay language writing: An exploratory study. *Pertanika Journal of Social Sciences and Humanities, 29*(3). https://doi.org/10.47836/pjssh.29.s3.11

Saralajew, S., Shaker, A., Xu, Z., Gashteovski, K., Kotnis, B., Ben-Rim, W., Quittek, J., & Lawrence, C. (2022). A human-centric assessment framework for AI. *ArXiv*, abs/2205.12749. https://doi.org/10.48550/arXiv.2205.12749

Selim, A. S. M. (2024). The transformative impact of AI-powered tools on academic writing: Perspectives of EFL university students. *International Journal of English Linguistics, 14*(1). https://doi.org/10.5539/ijel.v14n1p14

Sihombing, R. (2016). Teachers' problems and solutions in assessing students' writing in senior high school level: Authentic assessment or traditional assessment. In *Proceedings of the 2nd SULE – IC 2016, FKIP, Unsri, Palembang October 7th – 9th, 2016*.

Sireci, S. G., & Rizavi, S. M. (2000). Comparing computerized and human scoring of students' essays. *(ERIC Document Reproduction Service No. ED 463 324).*

Sokolov, C. (2014). Self-evaluation of rater bias in written composition assessment. *Lingüística, 54*, 261–275. https://doi.org/10.4312/linguistica.54.1.261-275

Son, J., Ružić, N., & Philpott, A. (2023). Artificial intelligence technologies and applications for language learning and teaching. *Journal of China Computer-Assisted Language Learning*.https://doi.org/10.1515/jccall-2023-0015

Sönmez, L., & Şeker, M. (2024). Investigating syntactic and morphological differences in the written productions of Turkish learners based on the learning context. *Kastamonu Education Journal, 32*(2), 272–281. https://doi.org/10.24106/kefdergi.1473587

Street, C. T. (2003). Pre-service teachers' attitudes about writing and learning to teach writing: Implications for teacher educators. *Teacher Education Quarterly, 30*, 33–50.

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*. https://doi.org/10.1016/j.asw.2023.100752

Sun, H. (2024). Multi-scenario application of ChatGPT-based language modeling for empowering English language teaching and learning. *Applied Mathematics and Nonlinear Sciences, 9*(1). https://doi.org/10.2478/amns-2024-0790

Torres, J. T. (2018). Constructing pre-service teacher identity within the discourse of writing assessment. *Practitioner Research in Higher Education, 12*(1).

Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. *Journal of China Computer-Assisted Language Learning, 3*, 258–262. https://doi.org/10.1515/jccall-2023-0008

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1). https://doi.org/10.1515/opis-2022-0158

Wang, R. E., & Demszky, D. (2023). Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the*

18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 626–667). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.53

Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023). Is ChatGPT a good NLG evaluator? A preliminary study. *ArXiv, abs/2303.04048*. https://doi.org/10.48550/arXiv.2303.04048

Wang, Z., & van Davier, A. A. (2014). Monitoring of scoring using the e-rater® automated scoring system and human raters on a writing test. *ETS Research Report Series, 2014*, 1–21. https://doi.org/10.1002/ets2.12005

Weiß, Z., Riemenschneider, A., Schröter, P., & Meurers, W. D. (2019). Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 30–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4404

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). Exploring the limits of ChatGPT for query or aspect-based text summarization. *ArXiv, abs/2302.08081*. https://doi.org/10.48550/arXiv.2302.08081

Yu, D. (2023). AI-empowered metaverse learning simulation technology application. In *2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA)* (pp. 1–6). https://doi.org/10.1109/iMETA59369.2023.10294830

Zhang, H., Jethani, N., Jones, S., Genes, N., Major, V. J., Jaffe, I. S., Cardillo, A. B., Heilenbach, N., Ali, N. F., Bonanni, L. J., Clayburn, A. J., Khera, Z., Sadler, E. C., Prasad, J., Schlacter, J., Liu, K., Silva, B., Montgomery, S., Kim, E. J., ... Razavian, N. (2024). Evaluating large language models in extracting cognitive exam dates and scores. *PLOS Digital Health, 3*(12), e0000685. https://doi.org/10.1371/journal.pdig.0000685